IMPUTATION IN AGRICULTURAL SURVEYS

H. F. Huddleston, U.S. Department of Agriculture
R. R. Hocking, Mississippi State University

1. Introduction

In sample surveys the respondents frequently do not cooperate and none of the information corresponding to the questionnaire items is collected. In other cases, the respondents provide only partial information because they are unwilling or unable to provide information for all items. In the first situation, all the questionnaire data is missing and a nonrespondent stratum is created. There have been several solutions proposed for missing data for a nonrespondent stratum irrespective of whether the nonresponse was due to not-at-homes or refusals. Hansen and Hurwitz, Politz and Simmons, and Hendricks have all proposed solutions depending on circumstances encountered. Hansen and Hurwitz propose a random sampling of the nonrespondent stratum while Politz and Simmons use the frequency that respondents may be found at home during the week as a weighting technique. Hendricks proposed relating the means in successive contacts with the respondents to determine a relation between the lack of willingness to cooperate (i.e., resistance) in order to estimate the mean for a characteristic. Each of these techniques is directed at estimating the mean or mean vector for a survey.

In the second situation, the questionnaire is incomplete, but it is possible to derive estimates from the full sample for some characteristics. In this case, the questionnaire can be treated as a multivariate response and interrelationships between characteristics may be examined to seek a solution using all the data. Also, by using appropriate survey strategy it may be possible to avoid a nonrespondent stratum and to treat the first situation as an incomplete multivariate response. This strategy can be employed in agricultural area sample surveys. A few characteristics, such as total land area in a farm tract, number of livestock and poultry visible, type of animal shelters or pens present, or cultivated acres in the farm tract may be determined through observation or photo measurements. Also, lists from State agricultural census may contain historical variables which are related to the questionnaire items which can be secured for all sampling units. More recently, crop acreage information for individual area sampling units has become available from satellite classification methods. By such means a few items may be available for the full sample and it is possible by employing interrelationships to estimate missing items. The success of this technique depends largely on the magnitude of the correlations between the questionnaire items which are available and those that are missing.

This paper describes a multivariate procedure which has been developed and tested on selected surveys over the past 10 years. The role of editing has been considered that of identifying the erroneous or missing entries in individual survey questionnaires. The role of imputation has been twofold: (1) to derive the best estimate of survey parameters, and (2) to impute values for the content items in individual questionnaires.

This paper focuses on a direct analysis of survey data using a technique based on a multivariate normal distribution and the principle of maximum likelihood which has been implemented by ESTMAT to achieve these goals. Only minor attention in the paper is devoted to techniques of editing or detecting unsatisfactory entries in questionnaires. All unsatisfactory entries, no matter how they arise, are considered as missing.

2. Estimating Parameters in Surveys with Incomplete Records

For analysis, the N records are divided into T groups, the t^{th} group containing n_t records which exhibit the same pattern of unsatisfactory entries. The missing data are assumed to occur at random in the sense that the joint distribution of observations is the product of the marginal distribution for the t^{th} group. An indicator matrix, D_t, is introduced to describe the data in the t^{th} group. Thus, if Z_{ti}, t=1, ..., T, i=1, ..., n_t, represents the p-vector which would have been observed when the record is complete, the vector Y_{ti} of length, say q_t, which is actually observed is given by

$$Y_{ti} = D_t Z_{ti} \tag{1}$$

For example, with p = 3 and Group 1 consisting of those records with the second entry missing, the indicator matrix for this group is

$$D_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

If, in Group 2, only the first entry and the sum of the second two entries are recorded, the indicator matrix is

$$D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \tag{3}$$

Note that to form D_t, we either eliminate or combine rows in the identity matrix in the same way that elements of Z_t are eliminated or combined to form Y_t.

The observed Z_{ti} are assumed to follow a p-variate normal, $N(\mu,\Sigma)$. Using this notation, the data in Group t as described by indicator matrix D_t of dimension q_t x p are q_t-variate normal, $N(\mu_t \Sigma_t)$ where

$$\mu_t = D_t \mu \tag{4}$$

and

$$\Sigma_t = D_t \Sigma D_t' \tag{5}$$

For the observations in Group t, let the sample mean vector and sample covariance matrix

be denoted by $\hat{\mu}_t$ and $\hat{\Sigma}_t$. Specifically,

$$\hat{\mu}_t = \sum_{i=1}^{n_t} Y_{ti}/n_t \tag{6}$$

and

$$\hat{\Sigma}_t = \sum_{i=1}^{n_t} (Y_{ti} - \hat{\mu}_t)(Y_{ti} - \hat{\mu}_t)'/n_t \tag{7}$$

The element of $\hat{\Sigma}_t$ in row j and column k is denoted by $\hat{\sigma}_{tjk}$ and is given by

$$\hat{\sigma}_{tjk} = \sum_{i=1}^{n_t} (y_{tij} - \hat{\mu}_{tj})(y_{tik} - \hat{\mu}_{tk})/n_t \tag{8}$$

where y_{tij} and $\hat{\mu}_{tj}$ denote the j^{th} components of Y_{ti} and $\hat{\mu}_t$.

It is convenient to display the elements of the symmetric matrix $\hat{\Sigma}_t$, of dimension q_t, as a vector of dimension $q_t(q_t + 1)/2$. This vector, denoted by $\hat{\sigma}_t$, has its components ordered according to the columns of $\hat{\Sigma}_t$, that is,

$$\hat{\sigma}_t = (\hat{\sigma}_{tjk},\ 1 \leq j \leq k = 1, \ldots, q_t) \tag{9}$$

The relation (5) between Σ_t and Σ may be expressed in vector form using an indicator matrix, C_t, of dimension $q_t(q_t + 1)/2 \times p(p + 1)/2$, which is constructed from D_t. Thus, we have

$$\sigma_t = c_t\sigma \tag{10}$$

The likelihood equations for estimating the parameters μ and σ using all of the N records are given by

$$W_\mu \mu = \sum_{t=1}^{T} D_t' W_{\mu_t} \hat{\mu}_t \tag{11}$$

and

$$W_\sigma \sigma = \sum_{t=1}^{T} C_t' W_{\sigma_t} (\hat{\sigma}_t + h_t) \tag{12}$$

Here W_{μ_t} and W_{σ_t} are the information matrices for μ_t and σ_t in the t^{th} group, and W_μ and W_σ are the overall information matrices, all depending on Σ. The vector h_t is formed as in (9) from the symmetric matrix

$H_t = (\hat{\mu}_t - \mu_t)(\hat{\mu}_t - \mu_t)'$ and hence depends on μ. For a development of (11) and (12) and further details, see Hartley and Hocking, 1971. In a recent technical report an alternate form of equation (12) is reported by Hocking (1977) which is more efficient to carry out computationally.

We illustrate the parameter estimation with several examples based on special purpose surveys or one section of a general purpose survey. We have not attempted to try the technique on a questionnaire with several hundred items without doing the imputation in small sections.

Example 1

The data for this example were complete and missing data were generated from a milk production survey conducted in Wisconsin in 1971. There is a total of N = 160 records with p = 5 variables.

There is a total of T = 10 groups of data as described in Table 1. In this table, an asterisk (*) indicates that the variable is observed in that group while a blank indicates that it is not.

Table 1. Description of Data for Example 1

Group no.	Group size	Variable				
		1	2	3	4	5
1	50	*	*	*	*	*
2	10	*		*	*	
3	10			*	*	
4	10	*	*	*		
5	10		*	*		
6	20		*		*	*
7	20		*	*	*	
8	10	*		*		
9	10	*	*			
10	10	*				
Total	160	100	110	120	110	70

The ESTMAT program was used to solve equations (11) and (12) using the sample covariance matrix from Group 1 to initialize the procedure. To provide comparisons, two other sets of estimates are provided in Tables 2 and 3. The first is obtained from the complete records from which the incomplete data were generated. The second set of estimates are obtained by a procedure often recommended for the analysis of such data. That is, to estimate μ, compute the sample means by pooling together those groups for which the item is recorded. Thus, the estimate of the first component of μ is based on the 100 observations in Groups 1, 2, 4, 8, 9 and 10. The estimates of the diagonal elements σ_{ii} of Σ are also obtained in the usual way from this univariate analysis. It is of interest to note that this procedure is identical with the maximum-likelihood procedure if, in fact, Σ is known to be diagonal but not otherwise. Thus, ESTMAT attempts to take advantage of the correlations between the variables to gain precision over the estimates obtained by the simple pooling procedure. The pooling procedure may also be used to provide an estimate of the σ_{ij}, $i \neq j$, but the properties of such an estimate are questionable. Alternatively, the sample covariance matrix for Group 1, $\hat{\Sigma}_1$ is often recommended.

Table 2. Estimates of the Mean Vector

Method	Variable				
	1	2	3	4	5
All data	12.96	14.79	18.87	13.76	5.78
ESTMAT	13.07	14.56	18.72	14.02	5.88
Pooled data	12.65	14.64	18.69	14.07	5.85

Table 3. Estimated Variances of Estimates of μ

Method	Variable				
	1	2	3	4	5
All data	0.176	0.128	0.129	0.181	0.013
ESTMAT	0.231	0.195	0.163	0.250	0.016
Pooled data	0.250	0.219	0.181	0.254	0.025

Comparison of the corresponding quantities for ESTMAT with those for the complete data provides an indication of the degree of precision lost due to the missing data. Comparison to the ESTMAT figures with those for the pooled data procedure indicates the relative efficiencies of the two methods. The variances derived using ESTMAT have the property of being greater than the variances which would have been obtained if all data were present but less than the variances for the "pooled estimator" when there are significant nonzero covariance terms.

Example 2

A livestock survey conducted in Texas in 1969 was concerned with an inventory of cattle. A total of $p = 4$ variables counting the number of cattle of various types was recorded. In addition, the total number of cattle of all types was recorded. Thus, it is possible to use a "total item" for estimating missing subgroup items.

If all records are complete and accurate, this fifth variable contains no additional information. In the case of the incomplete records, however, this variable does contain information and the purpose of this example is to illustrate how it can be used.

The ESTMAT procedure is capable of analyzing data which is incomplete because only linear combinations of certain variables were recorded. For example, some records may correctly record Y_1 and Y_2 and the total Y_5 but not give values for Y_3 and Y_4. We thus know the total of Y_3 and Y_4. Such data are described as Group 5 in Table 4. The appropriate indicator matrix for Group 5 is thus given by

$$D_5 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad (13)$$

For this example $T = 8$ groups of data were generated from $N = 160$ records which were judged to be complete and accurate. These groups are described in Table 4. In Table 4, an asterisk (*) indicates a variable observed, a blank indicates it is not observed and a plus (+) indicates that the sum of all variables with a (+) is observed. The number of observations in each group and the number of observations on each variable are also recorded.

Table 4. Description of Data for Example 2

Group no.	Group size	Variable			
		1	2	3	4
1	50	*	*	*	*
2	10		*	*	
3	10	*	*		
4	10	+	*	*	+
5	10	*	*	+	+
6	25	+	*	+	+
7	25	*	+	+	+
8	20	+	+	+	+
Total	160	120	120	110	70

If the information contained in Variable 5 (the total) is not available (or is not used), then, for example, Groups 2 and 4 have identical descriptions. For clarity, the description of the data in this situation is given in Table 5. In both cases, the sample covariance matrix from Group 1 was used to initialize the program and convergence was attained in three iterations.

Table 5. Description of Example 2 Ignoring Variable 5

Group no.	Group size	Variable			
		1	2	3	4
I (1)	50	*	*	*	*
II (2&4)	20		*	*	
III (3&5)	20	*	*		
IV (6)	25		*		
V (7)	25	*			
Total	140	95	115	70	50

An indication of the increased precision by using the information on the totals is obtained by comparing the estimated variances of the estimates of μ in Tables 6 and 7. A variable, such as an inventory total, which is available from the questionnaire, survey screening form, or historical file is helpful in estimating the parameters.

Table 6. ESTMAT Analysis of Table 4 Data

Estimate	Variable			
	1	2	3	4
$\hat{\mu}$	10.5	109.0	50.0	102.3
Variance of $\hat{\mu}$	20.2	105.6	80.1	124.4
σ_{ii}	2,056	15,074	6,877	12,963
Variance of $\hat{\sigma}_{ii}$ ($\times 10^{-3}$)	87.6	3,418	1,247	4,017

Table 7. ESTMAT Analysis of Table 5 Data

Estimate	Variable			
	1	2	3	4
$\hat{\mu}$	10.7	104.2	52.2	99.5
Variance of $\hat{\mu}$	22.6	123.0	111.2	200.6
σ_{ii}	2,184	14,276	7,837	11,774
Variance of $\hat{\sigma}_{ii}$ (x 10^{-3})	100.4	3,543	1,760	5,332

3. Procedures for Imputation for Individual Questionnaires

The procedures developed here are based on regression procedures employing the estimated survey parameters (mean vector and covariance matrix from last iteration of ESTMAT) derived in the preceding section as applied to the individual questionnaires in the groups prior to further analyses. That is, a single regression for each item for individual questionnaires in each group is derived from the estimated mean vector and covariance matrix for all groups. The imputed data are generally used where internal comparisons, or relations may be contemplated for subpopulations; or the individual units may be subsampled later for special purpose follow-up surveys.

We illustrate the procedure for imputing values for missing data from Group 9 of example 1 in which only Variables 1 and 2 are recorded. In this situation, the imputed values shown in Table 8 appear to be realistic only for about half the records, but probably acceptable for all because this variable is nonzero for all records. The missing variables are moderately correlated with the variables present.

A third survey data set is used to illustrate in more detail the imputation for individual questionnaires where a related linear combination of variables is employed.

Table 8. Individual Values for Variable 3 Group 9

Sample unit	Actual values	Regression imputed
1	378	375
2	525	380
3	457	338
4	168	346
5	322	339
6	426	415
7	164	325
8	284	372
9	308	311
10	360	372

Example 3

The data for this example were taken from one stratum of a pig survey conducted in Iowa in 1974. The number of items and sample sizes by group are shown in Table 9.

The asterisk (*) indicates the variable is observed, a blank indicates it is not observed, and a plus (+) indicates the sum of all variables with a (+) was observed. The groups were derived based on the completeness by subsections of the questionnaire. If the entries were considered questionable in a subsection, the whole subsection would be missing. This type of assumption is generally necessary in a long questionnaire in order to keep the number of groups to a manageable level for purposes of controlling matrix size and to insure an adequate sample size in each group to satisfy the requirement that $n_t > q_t$. The mean vector and its standard errors based on ESTMAT and the complete data set are given in Table 10.

The imputed values are examined only for items 1 and 7 in Groups 4 and 6. The predicted values are denoted as $\hat{y}_1$ for item 1 and the corresponding actual values as y_1. First, we examine Group 6 which is the nonrespondent stratum. This control variable represents a maximum derived number from the previous calendar year. The squared correlation coefficient between the individual items and the control variable ranged from .08 to .21. Consequently, it does not seem appropriate to consider (or show) imputed values for individual questionnaires. Imputed regression values for individual questionnaires under these conditions would seem to be feasible only if, say, the inventory total were known, or at least the inventory was known to be either zero or nonzero for the current data and used as a conditioning variable in deriving regression predictions for individual values.

For Group 4, the prediction of individual values are obtained from the following regression equation:

$$\hat{y}_{1i} = \hat{\bar{y}}_{1i} + \sum_{i=1}^{r} b_i(x_i - \bar{x}_i)$$

These values are shown in Table 11. The technique appears to work well for item 1, but is less satisfactory for item 7. This is a direct reflection of the magnitude of the squared multiple correlation coefficients which are .60 and .20, respectively. However, the imputation would be improved if the zero reports could be identified since they constitute a large fraction of the cases.

Table 9. Description of Example 3

Group no.	Group size	Variable 1/												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	166	*	*	*	*	*	*	*	*	*	*	*	*	*
2	12	*	*	*	*	*	*	*	*			*	*	*
3	12	*	*	*	*	*	*	*	*	*	*			
4	12	+	+	+	+	+	+	+	+	*	*	*	*	*
5	12	*	*	*						*	*	*	*	*
6	24													
Total	238	202	202	202	190	190	190	190	190	202	202	202	202	202

1/ One historical data variable was available for each questionnaire.

Table 10. Estimates of Means and Their Standard Errors

Subsection		Item no.	Item	Means		Standard errors	
				Reported	ESTMAT	Reported	ESTMAT
INVENTORY	TOTAL	1-8	Inventory	53.30	52.31	4.263	4.397
	BREEDING HERD	1	Sows	5.93	6.07	0.577	0.622
		2	Boars	.34	0.33	0.048	0.052
		3	Not breeding	.62	0.61	0.243	0.273
	MARKET HOGS BY WEIGHT	4	Pigs weighing under 60 lbs	15.83	15.01	1.984	2.081
		5	Pigs weighing 60-120 lbs	10.24	10.17	1.300	1.411
		6	Pigs weighing 120-180 lbs	12.54	11.67	2.045	1.860
		7	Pigs weighing 180-220 lbs	6.49	6.96	1.168	1.299
		8	Pigs weighing over 220 lbs	1.31	1.49	0.414	0.503
INTENTIONS		9	Expected farrowing first quarter	1.64	1.68	0.315	0.352
		10	Expected farrowing second quarter	3.98	4.02	0.452	0.480
MARKET RELATED		11	Farrowing last quarter	2.76	2.71	0.333	0.354
		12	On hand from all previous quarters	16.67	16.45	2.213	2.325
		13	Sold last quarter	1.34	1.46	0.636	0.734
CONTROL		14	Historical data-maximum size	103.39	103.39	3.414	3.414

Table 11. Regression Imputed and Actual Values for Items 1 and 7 Group 4 (Round to nearest integer)

Sample unit	Item 4		Item 7	
	$\hat{y}_1$	y_1	$\hat{y}_7$	y_7
1	4	0	12	0
2	27	24	36	75
3	4	4	7	0
4	0	0	-1	0
5	6	10	14	0
6	0	0	14	0
7	0	0	-1	0
8	16	12	2	9
9	0	0	-1	0
10	9	15	7	0
11	0	0	-1	0
12	3	0	8	3

4. Comments

The analysis of incomplete normal data according to the technique described in Section 2 is easily accomplished using the ESTMAT program. Experience with both simulated and actual data indicates that the iterative procedure does converge and that the convergence is rapid, commonly requiring less than three iterations depending on the structure of the incomplete data and the initial values used. Trials with simulated data indicate that the large sample-variance estimates are quite acceptable for moderately small samples.

The data in the examples presented in the paper were nonnormal in two respects. First, histograms of the marginals revealed that the distribution was skewed and second that there is a spike at the origin due to the high frequency of zeros. The skewness is easily removed by a square-root transformation, but the spike at the origin suggests a mixture of two distributions. If transformations to obtain normality are indicated, then they should, in general, be made. However, the estimates of the parameters, in the examples studied, were essentially equivalent whether the original data or the transformed data were used in the analysis indicating the parameter estimation is not highly dependent on the normality assumption. The problem of handling the spike at the origin is currently under investigation.

The estimation of survey parameters with incomplete questionnaires appears to be quite satisfactory for the population mean vector and its standard errors for the fraction of faulty records experienced (i.e., .10-.30) for questionnaires of modest length or subsections with related content items. The survey errors as stated are larger than if no data were missing but less than the pooled univariate estimate based on using only the items present. In addition, the estimates appear to be fairly robust under moderate departures from normality. If no survey data are available for a group, the earlier techniques of Hansen and Hurwitz, Politz and Simmons, or Hendricks would seem to be suitable for parameter estimation if time is available and circumstances are appropriate. Generally, a survey strategy which can employ several variables that are related to subsections of the questionnaire can be quite helpful in the estimation of survey parameters based on treating the questionnaire as a multivariate response. While ESTMAT will generally be satisfactory for parameter estimation, the resulting regression equations derived from the variance-covariance matrices for the strata (or population) may not be satisfactory for obtaining values for individual questionnaires.

A global survey question, such as total inventory of livestock by species or total cultivated area or crop land which is correlated moderately with subsections is necessary for imputation for individual questionnaires. Lacking this type of information, then the determination of whether the item or group of items are zero is quite helpful. The imputation for individual questionnaires is likely to be satisfactory only if there are related items present on the questionnaire with a square multiple correlation of at least .4. However, a single survey question whether it is reported currently or historically is unlikely to be satisfactory in predicting for more than a few items on the questionnaire.

If 30 percent or more of the questionnaires are expected to have missing entries, a survey strategy to insure related information for each sampling unit would appear worthwhile for imputation for individual questionnaires in agricultural surveys. This assumes the incomplete questionnaires are scattered over most of the subpopulations of interest rather than being concentrated in a single subpopulation. The basis for determining the number and composition of the T groups should be examined as a means of attempting to insure correlations among items and to keep the total number of groups to a manageable level.

REFERENCES

1. Hansen, M. H. and Hurwitz, W. N., "The Problem of Nonresponse in Sample Surveys," JASA, Vol. 41, 1946, pp. 517-529.

2. Politz, A. and Simmons, W., "An Attempt to Get the 'Not-At-Homes' Into the Sample Without Call-Backs," JASA, Vol. 44, 1959, pp. 9-31.

3. Hendricks, W. A., The Mathematical Theory of Sampling, The Scarecrow Press, New Brunswick, N.J., 1956.

4. Hartley, H. O. and Hocking, R. R., "The Analysis of Incomplete Data," Biometrics, 27, (1971), pp. 783-823.

5. Hocking, R. R., Technical Report on Design of Sample Surveys to Reduce Respondent Burden, (1977), Department of Computer Science and Statistics, Mississippi State University.